UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Transformer-Based Siamese Textual Encoder for Knowledge Base Completion

by

Mengyao Li

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Analytics (Reasarch)

Sydney, Australia

2020

Certificate of Authorship/Originality

I, Mengyao Li, declare that this thesis, is submitted in fulfilment of the requirements for the award of Master of Analytics (Research), in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

Signature:

Date: 23/02/2020

Dedication

To my parents Xiuhua Ren and Guanhai Li

Acknowledgements

First of all, many thanks to my principal supervisor Dr. Jing Jiang and co-supervisor Dr. Guodong Long. During chasing the degree of research master, my supervisors gave me careful guidance and kind care in my study, research, and life. Their profound knowledge, rigorous academic attitude, and tireless study spirit made me admire, and left an indelible impression on me, which greatly influenced and inspired me. I will bear their edification in my mind, and these will definitely benefit me for my whole life.

Secondly, I want to thank my group mates, Bo Wang, Tao Shen, Zonghan Wu, Yang Li, Hao Huang, Peng Zhang for their help in all aspects of the model implementing and thesis writing. The gave me tremendous help with the questions I raised and put forward many valuable suggestions for my codes and thesis, which benefit me a lot.

Finally, I am deeply grateful to my parents, Xiuhua Ren and Guanhai Li. Every progress I have made is condensed by their selfless dedication, support, and encouragement; thanks to every teacher who has taught me in the growth of the patient teaching; thanks to all my friends who mutually helped to grow up together and the friendship deserves my cherish. These lovely people will continue to encourage me to move forward for more solid and comprehensive works in the future.

Mengyao Li

Sydney, Australia, 2020.

List of Publications

Journal Papers

J-1. **Mengyao Li** and Jing Jiang, "Siamese Network for Knowledge Base Completion," *Neural Processing Letters*, pre-acceptance, 2020.

Contents

List of Figures

ABSTRACT

Transformer-Based Siamese Textual Encoder for Knowledge Base Completion

by

Mengyao Li

Knowledge graph is constructed as a directed graph and regarded as a kind of knowledge base. Its nodes and edges represent named entities and relations between the entities, respectively. Knowledge graph is normally denoted as a collection of triples, i.e., (head/subject entity, relation, tail/object entity), for feasible processing. However, incompleteness or sparsity issue usually exists in knowledge graph since the knowledge graph are constructed via human labeling. Therefore, *link prediction* task is proposed to complete a knowledge graph via predicting the missing links.

Previous approaches for link prediction on knowledge graph are mainly based on graph embedding, such as translation-based embedding approaches, convolutional-based embedding approaches, and graph convolutional network (GCN-based) approaches. These traditional graph embedding-based approaches straightforwardly learn the embeddings by considering a knowledge base's structure but are inherently vulnerable to the graph's sparseness or incompleteness issue. In contrast, some recent approaches, named textual encoding approaches here, learn the embeddings from a natural language processing perspective. That is, except for the structural information in a knowledge graph, the semantic information of the entities and relations, like their texts, mentions, and descriptions, are taken into account to enrich the embeddings of the knowledge graph. Hence, they capture such structured knowledge from the semantic perspective and employ deep neural text encoder to model graph triples in semantic space, but fail to trade off the contextual features or model's efficiency.

Therefore, in line with previous textual encoding approaches that represent graph components like entities or triples into latent semantic space, we propose a Siamese textual encoder operating on each graph triple from knowledge base, where the contextual features between a head/tail entity and a relation are well-captured to highlight relation-aware entity embedding while a Siamese structure is also adapted to avoid combinatorial explosion during inference of link prediction.

In the experiments, the proposed approach reaches the best or comparable performance on two link prediction datasets. And further compared to its baseline (i.e., a state-of-the-art textual encoding approach), our approach can accelerate the inference procedure by one or two orders of magnitude with an even better quality of predictions.

Abbreviation

KG - Knowledge Graph

KB - Knowledge Base

Conv - Convolution

CNN - Convolutional Neural Network

GCN - Graph Convolutional Network

R-GCN - Relational Graph Convolutional Network

NLP - natural language processing

LSTM - Long Short-Term Memory

GRU - Gated Recurrent Unit

RNN - Recurrent Neural Network

BPE - byte-pair-encoding

CBoW - Continuous bag-of-words

FFN - feed-forward network

SAN - self-attention network

LM - language model

MLM - masked language model

NSP - next sentence prediction

ELMo - Embeddings from Language Models

BERT - Bidirectional Encoder Representations from Transformers

RoBERTa - Robustly Optimized BERT Pretraining Approach

KG-BERT - Knowledge Graph BERT

GPT - Generative Pre-Training

SSP - semantic space projection

PMI - point-wise mutual information

Nomenclature and Notation

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
$\mathsf{\mathbf{A}}$	A tensor
$\boldsymbol{I}_n$	Identity matrix with n rows and n columns
$\boldsymbol{I}$	Identity matrix with dimensionality implied by context
$\boldsymbol{e}^{(i)}$	Standard basis vector $[0, \dots, 0, 1, 0, \dots, 0]$ with a 1 at position i
$\text{diag}(\boldsymbol{a})$	A square, diagonal matrix with diagonal entries given by $\boldsymbol{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{A} \backslash \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$

$\mathcal{G}$	A graph
$Pa_{\mathcal{G}}(\mathrm{x}_i)$	The parents of x_i in $\mathcal{G}$

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
a_{-i}	All elements of vector $\boldsymbol{a}$ except for element i
$A_{i,j}$	Element i, j of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,i}$	Column i of matrix $\boldsymbol{A}$
$\mathsf{A}_{i,j,k}$	Element (i, j, k) of a 3-D tensor $\mathbf{A}$
$\mathbf{A}_{:,:,i}$	2-D slice of a 3-D tensor
a_i	Element i of the random vector $\mathbf{a}$

Calculus

$\frac{dy}{dx}$	Derivative of y with respect to x
$\frac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\nabla_{\boldsymbol{x}} y$	Gradient of y with respect to $\boldsymbol{x}$
$\nabla_{\boldsymbol{X}} y$	Matrix derivatives of y with respect to $\boldsymbol{X}$
$\nabla_{\mathbf{X}} y$	Tensor containing derivatives of y with respect to $\mathbf{X}$
$\frac{\partial f}{\partial \boldsymbol{x}}$	Jacobian matrix $\boldsymbol{J} \in \mathbb{R}^{m \times n}$ of $f : \mathbb{R}^n \to \mathbb{R}^m$
$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	The Hessian matrix of f at input point $\boldsymbol{x}$
$\int f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral over the entire domain of $\boldsymbol{x}$
$\int_{\mathbb{S}} f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral with respect to $\boldsymbol{x}$ over the set $\mathbb{S}$

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$ or $\mathbb{E} f(x)$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$
$\mathrm{Var}(f(x))$	Variance of $f(x)$ under $P(\mathrm{x})$
$\mathrm{Cov}(f(x), g(x))$	Covariance of $f(x)$ and $g(x)$ under $P(\mathrm{x})$
$H(\mathrm{x})$	Shannon entropy of the random variable x

$D_{\mathrm{KL}}(P\|Q)$	Kullback-Leibler divergence of P and Q
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Functions

$f : \mathbb{A} \to \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$. (Sometimes we write $f(\boldsymbol{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log x$	Natural logarithm of x
$\sigma(x)$	Logistic sigmoid, $\dfrac{1}{1+\exp(-x)}$
$\zeta(x)$	Softplus, $\log(1+\exp(x))$
$\|\|\boldsymbol{x}\|\|_p$	L^p norm of $\boldsymbol{x}$
$\|\|\boldsymbol{x}\|\|$	L^2 norm of $\boldsymbol{x}$
x^+	Positive part of x, i.e., $\max(0, x)$
$\mathbf{1}_{\text{condition}}$	is 1 if the condition is true, 0 otherwise